Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/75/20240556

Improvement of the robustness of deep learning models against adversarial attacks

Yangfan Zhao

School of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China

yangfanzhao9@gmail.com

Abstract. Artificial intelligence has had significant and widespread advancements globally in recent years, primarily due to the exceptional progress of deep learning, a subset of machine learning. Currently, there is a rapid and significant rise in the adoption of deep learning as a research tool across several scientific disciplines worldwide. It has emerged as an integrative and versatile method for conducting general scientific research. Nevertheless, the extensive prevalence and ongoing refinement of adversarial assaults have become an intrinsic drawback of deep learning models, compromising the model's stability. Hence, enhancing the resilience of deep learning models against adversarial attacks has emerged as a highly focused area of research. This study employs the research method of literature analysis and literature review to examine the resilience and vulnerability of deep learning. It explores several approaches to enhance the robustness of deep learning models and provides a comprehensive summary and insights into improving these models.

Keywords: deep learning, robustness, adversarial attacks, model improvement.

1. Introduction

Deep learning models have shown significant success in image classification, speech recognition, and emotional interaction in recent years. Nevertheless, the model's robustness has been significantly challenged with the emergence of adversarial samples. Hence, it is imperative to use diverse elements, such as integration, distillation, and fine-tuning, to enhance the resilience of deep learning models.

Carsten Hartmann argues that the majority of applied learning models suffer from a lack of interpretability, and without knowledge of the model's specifics, potential robustness issues cannot be addressed. Hence, it is imperative to enhance the level of mathematics comprehension in order to enhance the reliability and robustness of the application. Furthermore, it illustrates that there are numerous constraints in the implementation of AI, but the potential of mathematical analysis might help to mitigate some of these constraints. If we acknowledge that the interpretable deep learning (DL) is not comprehended in the context of a scientifically interpreted deductive-jurisprudence model, then Bayesian probability theory can offer a method to interpret the DL in a precise statistical (abduction) sense. This approach effectively resolves the issue of overconfident prediction data that deviates from the training dataset.

This paper utilizes literature analysis and literature review as methods of research to examine the intrinsic properties of robustness and adversarial attacks in deep learning. Building upon this analysis,

DOI: 10.54254/2755-2721/75/20240556

the paper investigates approaches to enhance robustness. It offers partial guidance and concise recommendations for enhancing the performance of deep learning models.

2. Robustness of the Deep Learning Model

Deep learning is a subset of machine learning that specifically focuses on representation learning. It replicates the operational mechanism of the human nervous system and categorizes and isolates characteristics from data using multi-layer neural networks. Deep learning has the ability to automatically categorize data and derive more complex and abstract characteristics from it. The fundamental architecture of deep learning encompasses multi-layer perceptrons, convolutional neural networks, and recurrent neural networks.

The resilience of the deep learning model pertains to its ability to remain stable and reliable when confronted with adversarial attacks, as well as the presence of noise and interference in the input data. These characteristics encompass the inherent resistance to variations in the data distribution, as well as the lowest level of resistance to adversaries (sometimes referred to as adversarial robustness) [1]. A deep learning model with strong resilience may consistently achieve precise and stable learning outcomes even in the presence of increased interference and adversarial attacks.

Studies have demonstrated that deep learning models are susceptible to adversarial examples, which have minimal differences from genuine data yet pose challenges in their accurate identification. Indeed, numerous research have demonstrated that adversarial attacks represent an intrinsic vulnerability of deep learning models [2]. Therefore, we propose to summarize this phenomena as the susceptibility of deep learning models. To address this issue, we provide research directions to enhance the deep learning model's robustness from the standpoint of robust optimization, with the goal of enhancing the deep learning model's stability.

3. Adversarial attacks

In order to assess the resilience of the deep learning model, we employ adversarial attacks. The author chose three prevalent techniques for generating adversarial attacks: The Fast Gradient Sign Method (FGSM) is a straightforward and efficient technique for creating adversarial samples. It achieves this by introducing little perturbations to the input data in order to optimize the loss function. PGD, also known as Projected Gradient Descent, is an iterative technique used in adversarial training. It generates adversarial perturbations by repeatedly applying stochastic gradient descent (SGD) algorithms. The CW (Carlini-Wagner) assault is a commonly used adversarial attack technique that use optimization to minimize the loss function and generate smaller, less noticeable adversaries.

The study suggests that the confrontation sample exhibits the following qualities and characteristics. The first aspect is minuteness, which pertains to including an antagonistic approach that is either imperceptible to the human eye or, if apparent, does not significantly impact the final outcome. Adversarial resistance is the level of perplexity that an antagonistic sample causes to the original classifier. This can be quantified by assessing the accuracy of the classifier's classification. Robustness is a characteristic where the adversarial sample retains its capacity to attack the model even after undergoing complex processes such as illumination changes, deformation, denoising, conversion, or defense mechanisms [4]. Migrability, in this context, refers to the capacity to retain a consistent level of adversarial ability across several categorization models, even without direct access to the underlying model [3].

4. Methods to Improve the Robustness

4.1. Preprocessing Method

To enhance the accuracy and stability of the learning model, we employ a pre-processing technique to modify the input data to better suit the model's requirements. This involves performing specific operations and transformations on the data prior to feeding it into the model.

DOI: 10.54254/2755-2721/75/20240556

4.1.1. Data Augmentation. In order to enhance the generalization capability of deep learning models, it is necessary to accurately train them using a substantial volume of training data. Nevertheless, the process of gathering data can be particularly costly in certain trials, therefore necessitating the implementation of data augmentation techniques to mitigate this issue.

In the context of image classification, let's consider the application of enhancing the performance of an algorithm. After analyzing the enhancement techniques, enhancement rate, and the size of the basic dataset for each label [5], ResNet-20 and LeNet-5 were chosen as the experimental models for CIFAR-10 and MNIST datasets. The training effects of the basic model and the hybrid model (after applying data enhancement techniques) were thoroughly compared. As a result, 10 advanced data enhancement methods were proposed, including rotation, zoom, translation, inversion, solar energy, shear, histogram equalization, automatic contrast, color balance, and shear. By choosing the suitable enhancement rate (2-3 times) and training set, we may obtain a pretty optimal training model using data augmentation.

4.1.2. Regularization Method. The regularization method is designed to minimize testing errors rather than training errors by incorporating penalty terms into the loss function. This technique is used to improve the generalization of the model. It enables training even with small training sets or faulty optimization procedures, and may be easily applied to unseen data. Common regularization approaches encompass data regularization, model architecture regularization, error function regularization, regularization term regularization, and optimization algorithm regularization.

For instance, while considering model structure regularization [6], we must select properties or make assumptions that are special in order to achieve regularization. The relationship between input and output can exhibit a favorable property alignment with the dataset, or it can approximate the dataset by making simplified assumptions about the mapping. An example of a decision that affects the mapping process is determining the appropriate number of layers and cells. This decision is not straightforward, since it aims to prevent both underfitting and overfitting. Another example is the consideration of certain invariances in the map, such as the locality and displacement of feature extraction, which are fixed in the convolution layer.

4.2. Basic Ideas and Processes of Confrontation Training Methods

The principles and characteristics of adversarial training methods may differ, but the overall method can be categorized into three distinct phases.

Generate the adversarial sample: by adding the appropriate perturbation amount to the input sample, together with the loss function and gradient information, the adversarial sample is obtained. Training the adversarial model involves inputting the generated adversarial sample into the adversarial training model. This process leads to the calculation of fresh losses and the accumulation of the gradient.

Iterative update: Continuously repeat the aforementioned processes and iterate until the model reaches convergence or the predetermined stop condition is reached. The training model acquired by this approach has excellent resilience and capacity for generalization.

4.3. Model Improvement Method

4.3.1. Improvement of the Classifier. Optimization loss function: The loss function in machine learning can be broadly categorized into regression loss and classification loss. Based on this premise, the often employed measures of identification, detection, and segmentation can also be broadened. Choosing an appropriate loss function for model training can significantly enhance the optimization of classification problem performance.

Transfer learning refers to the process of applying knowledge gained from one task to another related task. The essence of transferring a trained model from one job to another is identifying the resemblance between the source domain and the target domain, and effectively leveraging this

DOI: 10.54254/2755-2721/75/20240556

resemblance. The concept of Fine-Tune is utilized to optimize the performance of classifiers in large-scale datasets by fine-tuning the existing network model in a complex training environment.

4.3.2. Improvement of the Model Structure. In recent years, researchers have increasingly concentrated on selecting or enhancing network architectures to effectively address specific problems and applications, as a result of the emergence of different network designs [7]. Therefore, it is necessary to investigate methods to enhance the architectural design of the model.

Firstly, adjust the depth and layer length of the network. Increasing the depth of the network enhances the model's learning and presentation capabilities by adding more layers. Widening the layer length involves increasing the number of neurons in each layer, which in turn increases the number of parameters and calculation complexity. This adjustment significantly improves the training speed of the model. Additionally, modifying the learning algorithm can significantly enhance the model's convergence speed and generalization capability. Different learning algorithms offer distinct advantages in this regard. Hence, it is imperative to select more effective algorithms based on various learning task models. The network design technique is employed to address the issue of choosing suitable network families and customizing the related structural modulation. It involves designing various combinations of network structures and modules to enhance the efficiency and performance of the model structure.

4.3.3. Integration and distillation methods. Integrated learning: Neural networks exhibit significant variability, causing the model's outcomes to be extremely sensitive to the starting parameters. This sensitivity makes it challenging to replicate the results. Nevertheless, implementing the integrated model allows for the consolidation of prediction outcomes through the training of many models, so effectively mitigating the network model's large variance. The typical approach involves employing several models, utilizing multiple neural networks with identical configurations, inputting the same training dataset, and randomly initializing distinct parameters to accomplish the goal of integrating neural networks. The training outcomes are frequently superior than those of any one model.

Distillation method: Knowledge distillation is a process that involves extracting the knowledge from a complicated model and transferring it to a simpler model. This can be done in various machine learning scenarios, resulting in the development of alternative frameworks [8]. Examples of techniques include model compression, which involves reducing the size of big discriminative models, compact predictive distributions in Bayesian inference, and the refinement of non-standardized generative models into tractable models.

5. Conclusion

This paper provides a summary of research on the resilience of deep learning models and adversarial attacks. It covers topics such as model resilience, methods for generating adversarial attacks, characteristics of adversarial samples, and techniques for improving model resilience. The information presented in this paper offers theoretical and methodological support for enhancing the resilience of deep learning models. Nevertheless, there are certain constraints that need to be addressed. These include the limited breadth of the research, which fails to encompass a complete range of topics, the omission of an introduction to all the methods employed, and the absence of relevant research data to substantiate the findings. In the future, it is imperative that we do a thorough analysis of the data training results in order to determine the most effective approach for improving the relevant experimental model. Within the realm of deep learning models, our objective is to develop a more all-encompassing theory that addresses the potential shortcomings of neural networks. This endeavor necessitates a deeper grasp of mathematical concepts, which is the current focus of our efforts.

References

[1] Fartash Faghri.Training Efficiency and Robustness in Deep Learning. Department of Computer Science, University of Toronto, 2022.

DOI: 10.54254/2755-2721/75/20240556

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks [J]. arXiv preprint arXiv:1706.06083,2019.

[3] Pan WW, Wang XY, Song ML, Chen C.Survey on generating adversarial examples.Ruan Jian Xue Bao/Journal of Software, 2020, 31(1):67-81 (in Chinese).http://www.jos.org.cn/1000-9825/5884.htm, 2023.

[4] Yan Z, Guo Y, Zhang C.Deep defense: Training DNNs with improved adversarial robustness [J].arXiv preprint arXiv:1803.00404, 2018.

[5] Benlin Hu, Cheng Lei, Dong Wang, Shu Zhang, Zhenyu Chen, A Preliminary Study on Data Augmentation of Deep Learning for Image Classification[J]. arXiv preprint arXiv:1906.11887, 2019.

[6] Jan Kukačka, Vladimir Golkov, Daniel Cremers, Regularization for Deep Learning: A Taxonomy[J].arXiv preprint arXiv:1710.10686,2017.

[7] Vamsi K Ithapu, Sathya N Ravi, Vikas Singh, On architectural choices in deep learning: From network structure to gradient convergence and parameter estimation[J]. arXiv preprint arXiv:1702.08670,2017.

[8] George Papamakarios, Distilling Model Knowledge[J]. arXiv preprint arXiv:1510.02437, 2015.